Journal of Biomolecular NMR (2019) 73:229–244
https://doi.org/10.1007/s10858-019-00251-7

ARTICLE

Validation of protein backbone structures calculated from NMR angular restraints using Rosetta

Joel Lapin[1] · Alexander A. Nevzorov[1]

Received: 24 February 2019 / Accepted: 2 May 2019 / Published online: 10 May 2019

Abstract

Multidimensional solid-state NMR spectra of oriented membrane proteins can be used to infer the backbone torsion angles and hence the overall protein fold by measuring dipolar couplings and chemical shift anisotropies, which depend on the orientation of each peptide plane with respect to the external magnetic field. However, multiple peptide plane orientations can be consistent with a given set of angular restraints. This ambiguity is further exacerbated by experimental uncertainty in obtaining and interpreting such restraints. The previously developed algorithms for structure calculations using angular restraints typically involve a sequential walkthrough along the backbone to find the torsion angles between the consecutive peptide plane orientations that are consistent with the experimental data. This method is sensitive to experimental uncertainty in interpreting the peak positions of as low as ± 10 Hz, often yielding high structural RMSDs for the calculated structures. Here we present a significantly improved version of the algorithm which includes the fitting of several peptide planes at once in order to prevent propagation of error along the backbone. In addition, a protocol has been devised for filtering the structural solutions using Rosetta scoring functions in order to find the structures that both fit the spectrum and satisfy bioinformatics restraints. The robustness of the new algorithm has been tested using synthetic angular restraints generated from the known structures for two proteins: a soluble protein 2gb1 (56 residues), chosen for its diverse secondary structure elements, i.e. an alpha-helix and two beta-sheets, and a membrane protein 4a2n, from which the first two transmembrane helices (having a total of 64 residues) have been used. Extensive simulations have been performed by varying the number of fitted planes, experimental error, and the number of NMR dimensions. It has been found that simultaneously fitting two peptide planes always shifted the distribution of the calculated structures toward lower structural RMSD values as compared to fitting a single torsion-angle pair. For each protein, irrespective of the simulation parameters, Rosetta was able to distinguish the most plausible structures, often having structural RMSDs lower than 2 Å with respect to the original structure. This study establishes a framework for de-novo protein structure prediction using a combination of solid-state NMR angular restraints and bioinformatics.

Keywords Oriented-sample NMR · Angular restraints · Dipolar couplings · Chemical shift anisotropy · Membrane proteins · Structure determination · Rosetta

Introduction

Over the last two decades, solid-state NMR (ssNMR) of uniaxially aligned samples has become a useful tool for structure determination of membrane proteins incorporated in planar, lipid-rich hydrated bilayers (McDonnell et al. 1993; Opella et al. 1999; Wang et al. 2001; Marassi and Opella 2003; Traaseth et al. 2006, 2009; Sharma et al. 2010; Verardi et al. 2011; Gayen et al. 2013; Gleason et al. 2013; Yamamoto et al. 2013). Here the structural information is directly obtainable from the positions of NMR resonances in the multi-dimensional spectra correlating the chemical shift anisotropy (CSA) of the protein backbone atoms (e.g. ^{15}N, ^{13}C), with the nearest-neighbor dipolar couplings (DC), such as ^{1}H–^{15}N (Wu et al. 1994) and ^{1}H$_{\alpha}$–^{13}C$_{\alpha}$ couplings (Sinha et al. 2007). For a protein aligned in a phospholipid bilayer, the principal axes of the chemical shift tensors and dipolar vectors have specific orientations with respect to the external magnetic field, B$_0$; thus, NMR frequencies become

✉ Alexander A. Nevzorov
alex_nevzorov@ncsu.edu

[1] Department of Chemistry, North Carolina State University, 2620 Yarbrough Drive, Raleigh, NC 27695-8204, USA

orientationally dependent. Ultimately, in order to calculate the overall protein fold, one has to relate these frequencies to the backbone ϕ/ψ torsional angles, which can then be used to reconstruct the polypeptide backbone of the membrane protein under study.

Peptide planes can be regarded as fixed-geometry units for calculating the orientational dependences of NMR resonances along the protein backbone. A peptide plane is outlined by four atoms in the backbone: an α-carbon, C_α^i, the amide proton H, the next α-carbon, C_α^{i+1}, and finally the carbonyl oxygen, O. Additionally, contained within the outlined atoms in the plane are the carbonyl carbon C’ of the i’th residue, and the nitrogen atom, N, which belongs to the $i+1$’th residue. Figure 1 shows three peptide planes, each one outlined in red. The orientation of the main magnetic field, B_0, relative to an arbitrarily chosen molecular frame associated with the peptide plane is given by two spherical angles: β, the longitudinal angle, and α, the azimuthal angle. These angles determine the measured frequencies for ^{15}N CSA, ^{1}H–^{15}N dipolar couplings, and any other in-plane resonances. By contrast, for the chiral $C_\alpha H_\alpha$ bond, which is located at the juncture of two adjacent peptide planes, its orientation with respect to B_0 additionally depends on the ϕ torsion angle between the two planes.

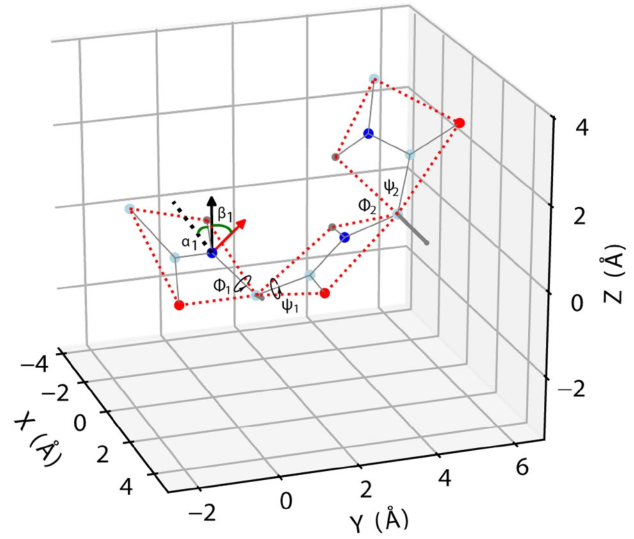


Fig. 1 An idealized representation for the backbone of a tripeptide unit. The edges of the peptide planes are outlined by the dotted red lines. The main atoms are color coded as: nitrogen in dark blue, carbon in light blue, oxygen in red, and hydrogen in gray. The orientation of the magnetic field B_0 relative to the molecular frame associated with the peptide plane is defined by two spherical angles β, α. The spherical angle β is measured from the external magnetic field, (black arrow) B_0, to the peptide plane normal (red arrow), which forms the z-axis for the molecular frame. The azimuthal angle α is measured between the NH bond, which forms the x-axis of the molecular frame, and the projection of B_0 onto the peptide plane (black dotted line). The relative orientations of the peptide planes are given by the pairs of torsion angles (ϕ_1, ψ_1) and (ϕ_2, ψ_2)

Structure determination from angular restraints relies on the measured frequencies to determine the possible peptide plane orientations along a polypeptide. In contrast to the methods employing distance restraints (Herrmann and Guntert 2002; Linge et al. 2003; Schwieters et al. 2003) and isotropic chemical shift data (Spera and Bax 1991; Cornilescu et al. 1999; Shen et al. 2009), structure determination of oriented membrane proteins from angular restraints still represents an outstanding problem. While it bears similarity to structure determination using RDCs (Valafar and Prestegard 2004; Bryson et al. 2008; Ruan et al. 2008), the number of observables per peptide planes is usually more limited and the spectra are measured relative to a single alignment frame defined by the external magnetic field B_0. Moreover, the high degree of correlation among the angular restraints renders global minimization methods such as simulated annealing impractical for de novo structure determination. In the mapping of resonance frequencies to peptide plane orientations, finding a consensus structure is primarily hindered by the degeneracy of the possible solutions, owing to the second-rank nature of the relevant chemical shift anisotropy and dipolar interactions. For instance, in the case of two experimental NMR dimensions, i.e. ^{15}N CSA and ^{1}H–^{15}N dipolar couplings there can be as many as 16 peptide plane orientations (pairs of β, α) consistent with each resonance (Bertram et al. 2003). Adding a third NMR dimension, namely ^{1}H$_\alpha$–^{13}C$_\alpha$ DC’s, significantly reduces the degeneracy (Yin and Nevzorov 2011). Including additional NMR spectroscopic dimensions can further help in resolving the correct structure from erroneous ones.

Another challenge for the structure determination process is experimental uncertainty, which can be thought of as ambiguity in interpreting the positions of the resonances. Such uncertainty can be due to insufficient resolution and slight variations in the dipolar scaling factors arising in composite decoupling pulse sequences (Wu et al. 1994; Dvinskikh et al. 2006; Nevzorov and Opella 2007). In addition, all resonances are fitted and then back-calculated assuming ideal peptide plane geometry and constant (average) chemical shift tensor principal values and orientations. In reality, minor deviations from both the ideal peptide plane geometry (Chellapa and Rose 2015) and the CSA tensor orientation and its principal values can arise depending on the amino acid type and local structure (Cornilescu and Bax 2000; Saito et al. 2010). This may interfere with the interpretation of angular anisotropy of the NMR resonances used during the structural fitting. For instance, it has been shown that in the limit of perfect resolution for each resonance, a 3D (^{15}N CSA, ^{1}H–^{15}N and ^{1}H$_\alpha$–^{13}C$_\alpha$ DC’s) spectrum yields a unique solution (Yin and Nevzorov 2011). When as little as 10 Hz of positional noise is added to the spectrum, the structural root-mean-square deviation (or structural RMSD) values may vary significantly.

A possible strategy to circumvent these problems is to develop a robust automated algorithm for finding a multitude of the possible structural solutions which could then be screened afterwards. One such previously developed structure fitting algorithm walks along the backbone and finds ϕ/ψ pairs that best-fit the adjacent peptide plane resonances (Yin and Nevzorov 2011). In contrast to independently determining the orientation of any two adjacent peptide planes and then solving a combinatorial search problem (Stewart et al. 1987, 1988; Bertram et al. 2003), this algorithm is recursive in the sense that the torsion-angle solutions for a given peptide plane pair are dependent on all ϕ/ψ solutions from the preceding residues. If at a certain residue the resonance frequencies cannot be fit within a desired tolerance, this may indicate implausible orientational solutions found for a previous residue(s). For such a scenario, a step-back or restart was implemented. In the presence of experimental uncertainty, the probability of misfit resonances increased greatly, even at a relatively low value of ± 10 Hz. Moreover, by trying to fit a single resonance at a time certain peaks may be over-fit, which may also result in erroneous ϕ/ψ pairs. In other words, the closest fit to the NMR resonance for the current residue may worsen the fit to the subsequent resonances, thus decreasing the average fit quality for the whole spectrum. Such erroneous ϕ/ψ pairs are often imperceptible by their fit to the spectrum alone and, thus, may go undetected. In the presence of experimental uncertainty, a plethora of possible structural solutions must be obtained in the hope that the solution set would still contain low-RMSD structures which could be distinguished from the rest.

It should be noted that structural solutions having large structural RMSDs (relative to the "correct" structure) could still yield good-quality fits to the spectrum. Nevertheless, while these structures may fit the experimental spectrum well, their overall conformations could violate basic protein folding principles and result in implausible topology. For example, the calculated torsion angles could occupy forbidden areas of the Ramachandran plot, or the backbone and side chains could overlap with each other or be too close in space. Moreover, there could be unstructured folds where compact secondary structure is expected, or the fold could be inconsistent with the protein's specific physical environment, such as hydrophobic side chains exposed to an aqueous environment, or vice versa, hydrophilic side chains contained in a hydrophobic lipid membrane interior. While such erroneous structures could violate any number of these conditions, the correct structure should satisfy them all. Bioinformatics can be of great use in applying physical and statistical criteria to discern the realistic solutions from the unrealistic ones. The Rosetta software package is a bioinformatic tool that has been very successful in de novo modeling of proteins (Kuhlman et al. 2003; Yarov-Yarovoy et al. 2005; Chaudhury et al. 2010; Leman et al. 2014). The popularity of Rosetta in structural biology is due to its ease of use and the comprehensive, well parameterized scoring functions. These scoring functions include numerous score terms that combine fundamental thermodynamic principles with statistical data compiled from real PDB structures. Moreover, the scoring terms can be made context-dependent, some specifically for soluble proteins and others for proteins immersed in a membrane environment. With the aid of Rosetta scoring terms, a post-fitting protocol can augment the NMR structural fitting algorithm to distinguish the reasonable structures from the ones that violate the aforementioned biophysical principles.

In the previously developed algorithm for finding the backbone structure (Yin and Nevzorov 2011), the search for the (ϕ_n, ψ_n) pair that best fit the NMR resonance for the n-th residue was carried out using a simplex-based minimizer. The simplex algorithm is convenient since it obviates explicit calculation of derivatives, which would be cumbersome given the recursive dependence of variables in the resonance fitting calculation. If the spectral root mean square deviation (or spectral RMSD) of the calculated frequencies from the NMR resonance is within a specified tolerance, the ϕ/ψ search is incremented to the next residue and the process continues all the way to the C-terminus of the backbone. If the tolerance is not satisfied for the nth residue fit, the algorithm steps back to the $(n-5)$'th residue and repeats that stretch of resonances in the hope of being able to find an acceptable solution on the next pass. In what follows, we describe a much improved and more robust algorithm which can be used for structure determination of macroscopically aligned proteins from their ssNMR spectra. Three major updates to the previously developed structure fitting algorithm have been implemented. The first update is to fix the number of simplex iterations during the search for the plausible torsion angles at each residue. The previous method would move forward to the next residue immediately after a pre-defined tolerance limit is satisfied. However, the initial selection of tolerances is arbitrary and, thus, the first acceptable solution below the tolerance might not be the best fit to a given spectral resonance. As a result of orientational degeneracy and experimental uncertainty there may be other torsion angles that would still score below the tolerance, even when using three or more NMR dimensions. Additionally, the simplex minimizer is highly dependent on the starting point, which requires numerous iterations to reliably find the global minimum. To keep searching even after the tolerance limit is satisfied makes it more likely that the best fit is found. The second modification is to search for the orientations of two (or more) planes at once, i.e. by fitting simultaneously the torsion angles ϕ_n, ψ_n, ϕ_{n+1}, ψ_{n+1}, etc. There will inevitably be some peptide planes where an erroneous ϕ/ψ

pair would best fit the current resonance while potentially resulting in poorly fit resonances downstream. Fitting the NMR resonances for several peptide planes simultaneously instead of a single resonance may help avoid over-fitting single resonances, thus avoiding specious ϕ/ψ solutions. The third major addition includes post-fitting filtering of the calculated structures using various Rosetta scoring functions to obtain consensus structural solutions.

These modifications yield a higher percentage of acceptable, low structural RMSD solutions when a multitude of structural fits to the data are possible. A protocol employing Rosetta scoring functions is presented, which accurately distinguishes the "true" structures from those that violate physical principles of protein folding and result in unlikely conformations while still satisfying NMR angular restraints. Since at present experimental solid-state spectra correlating three independent spectroscopic dimensions are not routinely available, the protocol was exemplified by using synthetic spectra. The method was tested on synthetic datasets generated from a soluble protein (PDB 2gb1) and the first two transmembrane helices of a membrane protein (PDB 4a2n). Structural solutions were obtained at various degrees of experimental error or uncertainty (also termed as "noise") in each NMR dimension. Experimental error was simulated by randomizing the NMR resonances, calculated from the known three-dimensional structures, in order to illustrate the performance of the algorithm in the presence of non-ideal peptide plane geometry, variable chemical shift tensor orientations, and uncertainty in the peak positions arising from spectral resolution. Rosetta scoring was used to determine the consensus structures for these proteins which were then compared to the original structures used to generate the NMR angular restraints.

Analytical framework: calculating protein structures in the spherical basis

For most angular-dependent NMR observables the orientation of the magnetic field B_0 relative to each peptide plane (molecular frame) is most efficiently represented in the irreducible spherical basis (Yin and Nevzorov 2011). As was previously mentioned, two spherical angles, β and α, specify the orientation of B_0 relative to a molecular frame associated with a given peptide plane. The three-dimensional irreducible row vector $\vec{Y}$ can be constructed using the following form:

$$\vec{Y}(\beta,\alpha) = \left(-\frac{\sin\beta}{\sqrt{2}}e^{i\alpha}, \cos\beta, \frac{\sin\beta}{\sqrt{2}}e^{-i\alpha}\right) \tag{1}$$

To convert the orientation vector into a scalar NMR observable, such as CSA or DC, we invoke an interaction tensor, M, which must be additionally transformed from the molecular frame (M) into its principal axis system (P). This transformation can be expressed by the Wigner rotation matrix $D(\Omega_{\mathrm{MP}})$, and any angular-dependent NMR observable can be written in the following generic form:

$$\nu = \vec{Y}(\beta,\alpha) \cdot \left[D(\Omega_{\mathrm{MP}}) \cdot M \cdot D^{+}(\Omega_{\mathrm{MP}})\right] \cdot \vec{Y}^{+}(\beta,\alpha) \tag{2}$$

Here the superscript "+" denotes the Hermitian conjugate. To propagate the orientation vector $\vec{Y}$ from residue n to residue $n+1$, a propagator matrix is applied:

$$\vec{Y}(\beta_{n+1},\alpha_{n+1}) = \vec{Y}(\beta_n,\alpha_n)P(\phi_n,\psi_n,\omega_n) \tag{3}$$

The propagator consists of the product of three Wigner rotation matrices, which explicitly contains fixed and variable angular parameters along the protein backbone (Yin and Nevzorov 2011):

$$\begin{aligned}P(\phi_n,\psi_n,\omega_n) = D(\alpha_{NC_\alpha},\phi_n,\gamma_{tetra})D(0,-\psi_n-180°,0)\\ D(-\alpha_{NC'C_\alpha},180°-\omega_n,-90°-\gamma_{HNC'})\end{aligned} \tag{4}$$

Here the angles $\alpha_{NC_\alpha} = 151.8°, \gamma_{tetra} = 110.5°, \alpha_{NC'C_\alpha} = 115.6°$, $\gamma_{HNC'} = 119.5°$, and $\omega = 180°$, can be treated as constants assuming an ideal peptide plane geometry. Calculating the $C_\alpha H_\alpha$ DC requires a different transformation, viz.:

$$\overrightarrow{Y_{CH}} = \overrightarrow{Y_n} \cdot D(270° - \alpha_{NC_\alpha}, \phi_n - 60°, 90° - \gamma_{tetra}) \cdot D(0,-90°,0) \tag{5}$$

From which the $C_\alpha H_\alpha$ DC can be explicitly calculated from the second element of the vector $\overrightarrow{Y_{CH}}(2)$ as:

$$\nu_{CH} = \chi_{CH}\frac{3\left[\overrightarrow{Y_{CH}}(2)\right]^2 - 1}{2} \tag{6}$$

where χ_{CH} is the DC constant for the $C_\alpha H_\alpha$ coupling interaction (see below). Assuming a constant peptide plane geometry, only two variables are contained in each propagator, namely the dihedral angles ϕ_n, ψ_n. Equations (1–5) represent a recursive recipe for mapping the backbone torsions ϕ/ψ onto the NMR observables, each specified by its own interaction tensor M.

The CSA interaction tensor for ^{15}N depends on its three principal components $\sigma_{11}, \sigma_{22}, \sigma_{33}$. Although the values of the principal components vary from residue to residue in real proteins, average tensor values are used in the calculations, and are written in the irreducible basis as:

$$M = \begin{pmatrix} 0.5*(\sigma_{11}+\sigma_{22}) & 0 & 0.5*(\sigma_{22}-\sigma_{11}) \\ 0 & \sigma_{33} & 0 \\ 0.5*(\sigma_{22}-\sigma_{11}) & 0 & 0.5*(\sigma_{11}+\sigma_{22}) \end{pmatrix} \tag{7}$$

For Glycine these values (in ppm) are: (41, 64, 215); and (64, 77, 222) for all other residue types. For the transformation of the matrix into the principal axis of the tensor, $D(\Omega_{MP})$, the σ_{33} axis of the CSA tensor was assumed to be first rotated by 18.5° off the NH bond within the peptide plane and then tilted by 25° off the plane normal about the σ_{22} axis. These values have been established experimentally in both solid-state and solution NMR (Lee et al. 1998; Cornilescu and Bax 2000). Possible variability in the tensor orientation along the protein is considered in the results and discussion sections. The DC interaction tensors differ only by their coupling constants, χ, each having the same form for the inner matrix M:

$$M = \chi\begin{pmatrix} 0.25 & 0 & -0.75 \\ 0 & -0.5 & 0 \\ -0.75 & 0 & 0.25 \end{pmatrix} \tag{8}$$

where the coupling constant χ between any two spins 1 and 2 is given by:

$$\chi = \frac{\mu_0\gamma_1\gamma_2 h}{16\pi^3 r_{12}^3}(Hz) \tag{9}$$

Here γ_1, γ_2 are the gyromagnetic ratios of the two interacting spins, and r_{12} is the distance between the spins. For ^{1}H–^{15}N the coupling constant is 9965.4 Hz, and for $C_\alpha H_\alpha$ it is 23,334.7 Hz.

In order to make the matrix–matrix multiplications more efficient, the quantities of Eqs. (2–5) can be transformed from the *Y* into the *Q*-basis, which diagonalizes the variable part for the rotations involving ϕ, ψ, and ω. The *Q*-transformations can be carried out as follows:

$$T_Q = e^{-i\left(3\pi/2-\alpha_{NC'C_\alpha}-\gamma_{HNC'}\right)L_y} \cdot e^{i\pi/2L_x} \tag{10}$$

$$M_Q = T_Q \cdot M \cdot T_Q^+ \tag{11}$$

yielding the following relations between the "*Y*" and "*Q*" bases:

$$\vec{Q} = \vec{Y} \cdot e^{-i\left(3\pi/2-\alpha_{HNC_\alpha}\right)L_y} \cdot e^{-i\pi/2L_x} \tag{12}$$

Here the matrices L_x, L_y are given by the rank-1 angular momentum operators:

$$L_x = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad L_y = \frac{i}{\sqrt{2}}\begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \tag{13}$$

Since the operations for calculating NMR observables are performed most frequently during the execution of the program, transforming the *Y*-basis into the diagonal *Q*-basis results in a 2-fold improvement in runtime. The *Q*-basis can be transformed back into the *Y*-basis using the inverse matrices corresponding to the operations given by Eqs. (10–12).

Methods

Structure fitting algorithm flowchart

All computer codes and scripts can be provided upon request to the authors. The algorithm for structure calculations has been implemented in the C programming language, which significantly speeds up the calculations versus the previous code (Yin and Nevzorov 2011) that was written in Matlab (Mathworks®). It should be noted that implementing the new algorithm in an interpreted computer language becomes prohibitively slow due to the sampling demands of the simplex solver while simultaneously fitting two or more peptide planes. The algorithm flowchart is displayed in Fig. 2, having the following basic steps:

(1) The input spectrum consists of a list of NMR resonance targets. A spectrum can either be simulated (option 1A) or read in (option 1B) from a file that includes the starting orientation, β_1, α_1, and the ϕ/ψ torsion angles for every residue.

(2) The master loop administers the calculation of kb_{max} structures, which continues until the last structure is reached, i.e. $kb = kb_{max}$.

(3) For each calculation, "experimental uncertainty" or "noise" can be included in the resonance targets by adding a random number, chosen from a uniform distribution between –$noise_{max}$ and +$noise_{max}$, to each dimension in the spectrum. This effectively puts the resonance to be fitted inside a cube (for three spectral dimensions) with side lengths equal to 2*$noise_{max}$. Simulated noise should be utilized on any spectrum, synthetic or real, in which there is uncertainty assumed in the peak center positions. All tolerances must be reset since they may have been altered in the previous run (see **8N,Y**). Due to the random noise added at the beginning, all alterations to tolerances only apply to that run.

(4) Enter loop for calculating structure. Start at first residue, $N = 0$.

(5) Is the current residue the first peptide plane, i.e. $N = 0$?

(6) If at the first residue (**6Y**), search for the starting orientation (angles β, α) of the magnetic field relative to the first peptide plane. Then convert the *Y*-basis to the *Q*-basis. Proceed to the torsion angle search (**6N**). If not at the beginning of the sequence then bypass **6Y** and proceed directly to **6N**. Set the tolerance for the current residue. Search for plausible ϕ/ψ combina-

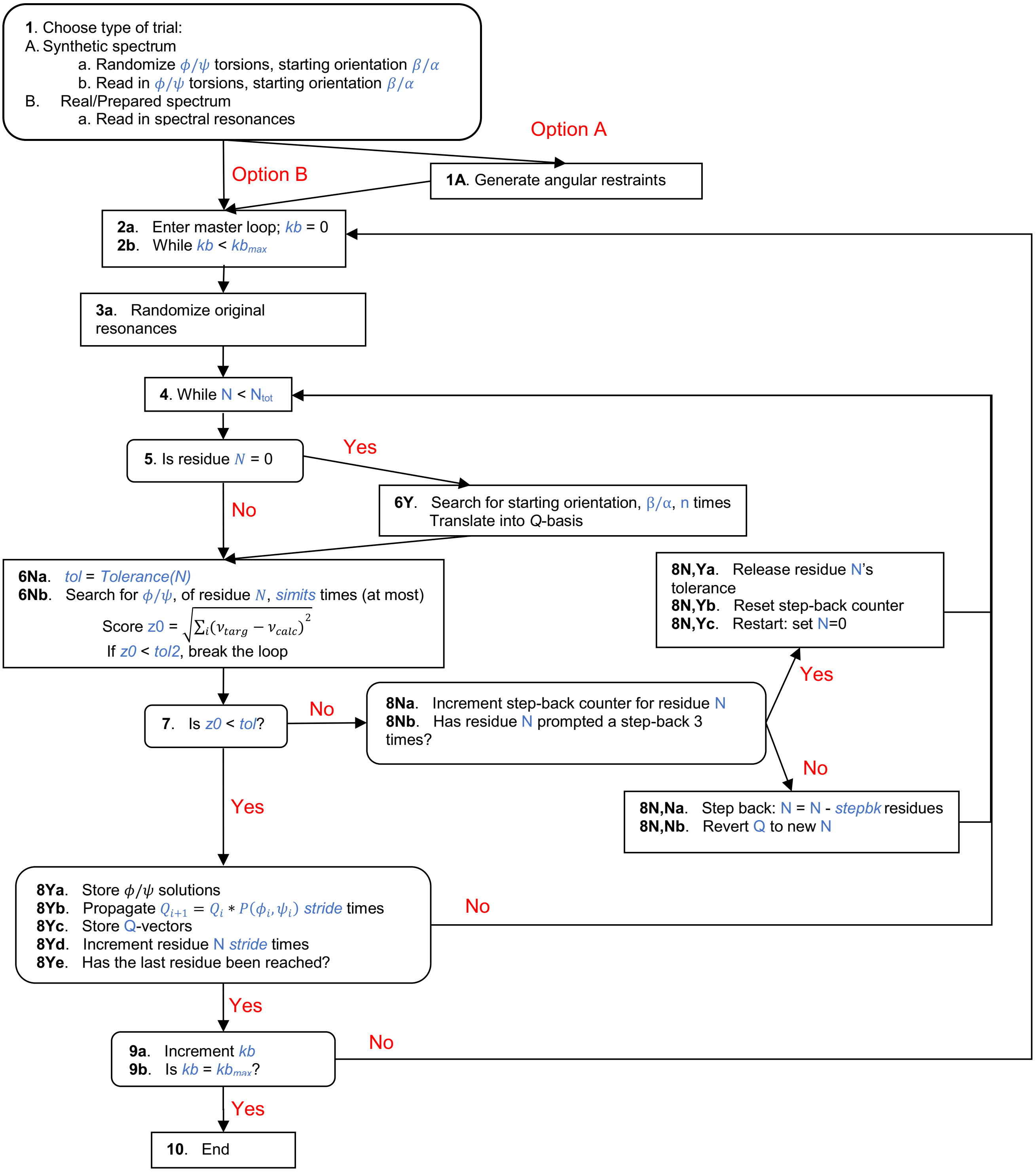

1. Choose type of trial:
A. Synthetic spectrum
a. Randomize ϕ/ψ torsions, starting orientation β/α
b. Read in ϕ/ψ torsions, starting orientation β/α
B. Real/Prepared spectrum
a. Read in spectral resonances
Option A
Option B
1A. Generate angular restraints
2a. Enter master loop; kb = 0
2b. While kb < kb_{max}
3a. Randomize original resonances
4. While N < N_{tot}
5. Is residue N = 0
Yes
No
6Y. Search for starting orientation, β/α, n times
Translate into Q-basis
6Na. tol = Tolerance(N)
6Nb. Search for ϕ/ψ, of residue N, simits times (at most)
Score z0 = $\sqrt{\sum_i (\nu_{targ} - \nu_{calc})^2}$
If z0 < tol2, break the loop
7. Is z0 < tol?
No
Yes
8Na. Increment step-back counter for residue N
8Nb. Has residue N prompted a step-back 3 times?
Yes
No
8N,Ya. Release residue N's tolerance
8N,Yb. Reset step-back counter
8N,Yc. Restart: set N=0
8N,Na. Step back: N = N - stepbk residues
8N,Nb. Revert Q to new N
8Ya. Store ϕ/ψ solutions
8Yb. Propagate $Q_{i+1} = Q_i * P(\phi_i, \psi_i)$ stride times
8Yc. Store Q-vectors
8Yd. Increment residue N stride times
8Ye. Has the last residue been reached?
No
Yes
9a. Increment kb
9b. Is kb = kb_{max}?
No
Yes
10. End

Fig. 2 Flowchart for the structure calculation algorithm. Variables in the program are represented with blue text

tions using a simplex search function in a loop that runs *simits* iterations. The number of torsion angles to be searched at once will be the 2**stride*; *stride* is the number of planes searched at once. Score the calculated resonances to the target spectrum (*z0*). To speed up the program, break the simplex loop if *z0* is less than *tol2* (optional).

(7) Is the score to the spectrum, *z0,* below the tolerance for residue *N*?

(8) If $z0 < tol$ (**8Y**) then store the torsion angles, propagate the orientation vector Q from residue N to residue $N + stride$, and increment N. Store the Q-vectors. If no $z0$ was found below the tolerance (**8N**) increment the step-back counter for residue N. If residue N fails to pass its tolerance 3 times (**8N,Y**), release its tolerance, reset the step-back counter, and step-back to the first residue. Otherwise (**8N,N**) step-back to residue *N-stepbk* and revert Q_N to $Q_{N\text{-}stepbk}$. If residue N is the last in the sequence, proceed to **9;** otherwise, go back to **4**.

(9) Structure calculation for structure *kb* is complete. Store the starting orientation and torsions. Increment *kb*. If $kb = kb_{max}$, exit the main loop and proceed to **10**. Otherwise go back to **2**.

(10) Finish the program. Print any pertinent information to the output files (torsion angles, atomic coordinates in the PDB format).

The quality of the output structures will depend on the size of the protein, the number of planes being evaluated per iteration, and the noise level being added to the spectral targets. The number of planes that are fit simultaneously is set by the variable *stride*. Since the simplex function is the bottleneck for program runtime, *simits* should not exceed an amount necessary to reliably find the lowest ϕ/ψ solutions for the majority of the runtime. With a larger value for *stride*, more simplex iterations will be necessary to ensure the lowest solution is found for every residue. The number of residues in the step-back move is set by *stepbk*, which was always set to 5 for these calculations. The experimental uncertainty level is set by *noisiness* and corresponds to the maximum deviation (in Hz) from the "true" spectral target in each dimension. The higher *stride* and *noisiness* are, the higher the tolerance criteria, *tol*, should be set. The program can be sped up by setting *tol2* at a value that below which there can be confidence that the simplex has already found the lowest function value for the current residue. Determining appropriate values for *tol* and *tol2* requires trial and error with the program. The general procedure used to find the right values involves setting *tol* high and *tol2* low, and gradually increasing and decreasing the two values, respectively, up to the point that they start compromising the quality of the structural solutions. Finally the number of structural solutions can be set by *kb_max*.

For the calculation of synthetic spectra, input csv files were prepared with all relevant information. A script was written in python, *torsions.py*, to read in the coordinates of a PDB file and calculate the ϕ/ψ angles for every residue. The ϕ/ψ torsions were written to a csv file whose top line was a single integer for the number of residues in the protein, the second line contained the one-letter amino acid codes for every residue, and the third and fourth lines contained the list of ϕ and ψ torsion angles, respectively. With the backbone torsion angles and a specified pair of angles (β, α) for the starting orientation, the spectrum was calculated using Eqs. 1–6 assuming ideal peptide plane geometry. Angles (β, α) were selected for membrane protein *4a2n* so that the helices spanning the membrane would be on average oriented along the z-axis, which represents the most biologically realistic orientation for transmembrane helices. There were multiple outputs for the program. The specific angles and parameters used in the calculations of the peptide plane, from Eq. 4–9, were written to a csv file named *angles.csv*, from which the calculations could be reproduced in post-fitting evaluation of the results. The results of the structure calculations were written to a file named *output.csv*. The structure of the csv file started with a comment line at the top, which included specific details for the simulation. The second line contained 2 integers, the first of which was the number of residues in the protein being measured, and the second was the total number of structural solutions calculated during the program run (*kb_max*). The third line was the one-letter amino acid sequence. All remaining lines are successive structural solutions that include 3 lines per solution: the first line is the (β, α) starting orientation solution obtained for that particular structure, and the second and third lines are the lists of ϕ and ψ torsions, respectively. For example, if a program run collected 1000 solutions then the output file would contain 3003 lines total. Finally the program writes another output in the same structure as *output.csv* containing only the orientation and ϕ/ψ torsions of the real system from which the spectrum was generated, named *real.csv*.

Program runtime largely depended on the value for *simits* used in the simulation. With the simplex function being the execution bottleneck in the program, running with $stride = 1$ was roughly an order of magnitude faster than $stride = 2$. For single-plane fitting the runtime per structural solution on a desktop computer equipped with i7 2 GHz Intel™ processor ranged between 5 and 30 s, largely depending on how many step-back moves were made for a particular solution. For fitting two planes at once the runtime was 60–220 s per solution.

Calculations of spectral and structural RMSDs

The figure of merit for spectral fitting is the magnitude of deviations from the target resonances, which can be calculated for a set of kb_{max} solutions as:

$$\overline{\Delta v_k^2} = \frac{\sum_j^{kb_{max}} \sum_i^N \frac{\left(|C_{i,k} - T_{i,k}|\right)^2}{N}}{kb_{max}} \tag{14}$$

The results of Eq. 14, $\overline{\Delta v_k^2}$, for each NMR dimension are then used to calculate the spectral RMSD of the fit to the spectrum (in Hz) as:

$$\Delta X = \sqrt{\sum_{k}^{M} \overline{\Delta v_k^2}(\mathrm{Hz})} \tag{15}$$

Here $C_{i,k}$ and $T_{i,k}$ are the calculated and target values for M-dimensional resonance, i in an N-residue protein in the k'th spectroscopic dimension. The average is taken over all calculated structures. In Eq. 15 the magnitude is taken over the M dimensions of the NMR spectrum being fit. In this work kb_{max} was always 1000 and M is 3 (corresponding to ^{15}N CSA, and ^{1}H–^{15}N and ^{1}H$_{\alpha}$–^{13}C$_{\alpha}$ dipolar couplings).

Structural RMSD calculations have been carried out using the python script *plotstruct.py*. This script reads in the orientation of the first peptide plane and the torsion angles, and calculates the N-residue structure into an (3N) × 3 array of points for atoms along the backbone. The script then reads in the torsions from *real.csv*, generates the "true" (PDB) backbone structure coordinates in the same size array and calculates the structural RMSD using the Kabsch algorithm (Kabsch 1976). Structural RMSD calculations are composed of a rigid translation to bring each structure's center of mass into coincidence, and a rigid rotation to minimize the distance of all corresponding points in two structures.

Rosetta post-fit filtering of structures

PyRosetta was used in order to utilize the Rosetta scoring functions. Inside the python scripts *score_output.py*, *membrane_score_output.py* the protein was assembled in a Rosetta *pose* object, using the output ϕ/ψ torsions from the program. The *pose* object was then scored in both full atom and centroid (coarse grained side chain representation) forms. The full atom function used for both soluble and membrane proteins was *talaris2013* (Kuhlman and Baker 2000; Kuhlman et al. 2003; Rohl et al. 2004; Leaver-Fay et al. 2013; O'Meara et al. 2015). The centroid function used for soluble proteins was *score3* (Rohl et al. 2004), and for membrane proteins it was *mpframework_cen_2006* (Alford et al. 2015). In order to further customize score functions, the values for the individual terms of each score function were written to file for all outputs from the original program. The python script *rosetta_scores.py* read in the unweighted individual terms and scores and recombined them with custom weights, effectively creating new custom scoring functions. The program also was used to scan different combinations of score terms in order to develop the most effective scoring function.

PyRosetta was used to distinguish the low-RMSD structures from the high ones. This involved multiple custom scoring functions that filtered results in two steps. The first filtering tier was a coarse search from which the top 20 solutions were extracted. The second tier applied a slightly different scoring function to sort out and rank those top 20 solutions, after which the top 10 solutions were reported as the final answer. The scoring functions were customized by finding the scoring terms, both full atom and centroid, that best correlated with structural RMSD in the two proteins being tested. Combinations of the scoring terms were scanned for those that produced the lowest median structural RMSDs in their top solutions. Ultimately the consensus scoring functions, i.e. the terms and their weights, were determined based on those that produced reasonable final structural RMSDs at all noise levels.

Table 1 lists the terms of the 2 scoring functions for soluble proteins, denoted with the letter 's'. Each scoring function is constructed by summing up the individual terms with the relative weights as given. These scoring terms are specific to soluble proteins (Leaver-Fay et al. 2013; O'Meara et al. 2015) and thus cannot be used for proteins in a membrane environment (although most terms have analogs for membrane proteins and vice versa). Analysis on 2gb1 solutions from the program found that term *rg* had the best structural RMSD correlation amongst all the structural solutions. This represents a statistical term favoring compact structures, i.e. having minimal radius of gyration. Terms *env* and *cbeta* also favor compact structures, and had the next best correlations. Term *env* is a solvation term for every residue based on their hydrophobicity, and *cbeta* is a solvation term correcting for excluded volume. Score function "1 s" contains all centroid terms, whereas score function "2 s" only contains full-atom terms. Term *fa_rep* is the Lennard-Jones repulsion energy between pairs of atoms, and *fa_solv* is the Lazaridis-Karplus solvation energy. The two scoring functions used for membrane proteins are denoted with the letter 'm' in Table 1. Term *mp_env* is a statistical term describing the likelihood of a specific residue being at the calculated depth in the modeled membrane; *rama* refers to the backbone torsion-angle preferences by residue type on the Ramachandran map. Once again, the second scoring

Table 1 Custom Rosetta score functions for soluble (s) and membrane (m) proteins

Score function	Term	Weight
1 s	rg	1.0
1 s	env	1.0
1 s	cbeta	1.0
2 s	fa_atr	0.2
2 s	fa_sol	1.0
1 m	mp_env	1.0
1 m	Rama	0.5
2 m	fa_atr	0.2
2 m	hbond_sr_bb	1.0
2 m	p_aa_pp	1.0
2 m	fa_mpsolv	1.0

function consists of terms that score full-atom structures and contain a mix of physical and statistical criteria. Term *fa_atr* refers to the Lennard-Jones attractive energy; *hbond_sr_bb* scores the hydrogen bonding between backbone atoms; *p_aa_pp*, similar to *rama*, is the probability for a given amino acid to have a given ϕ/ψ torsion pair; *fa_mpsolv* scores pairs of residues based on their depth in lipid bilayer. Scoring terms *mp_env* and *fa_mpsolv* are specific to membrane environments and available from Rosetta's membrane modeling package *RosettaMP* (Leman et al. 2014; Alford et al. 2015).

Results

Fitting of synthetic data for GB1 protein

All 55 residues of the soluble GB1 protein (PDB ID 2gb1; Gronenborn et al. 1991) were used for generating the spectrum consisting of ^{15}N CSA, and ^{1}H–^{15}N and ^{1}H$_{\alpha}$–^{13}C$_{\alpha}$ dipolar couplings. The first peptide plane orientation was arbitrarily set to $(\beta, \alpha) = (0.5, 1.0)$. There were 1000 structural solutions collected at experimental uncertainty ("noise") levels of 10, 30, and 50 Hz and *stride* number (either 1 or 2 planes). For the three noise levels *tol* was 50, 70, and 90 Hz respectively, and *tol2* was always 20 Hz. When *stride* = 1 (single-plane fitting), *simits* was set to 100, and increased to 1000 when *stride* = 2 (two-plane fitting). The structural RMSDs of 1000 back-calculated structures were binned as histograms shown in Fig. 3. The structural RMSDs were calculated relative to a "true" structure having the torsions angles calculated from the PDB coordinates. For each experimental uncertainty the 1000 structural solutions were put through the Rosetta post-fitting protocol as described in the Methods section. Using the 1000 sets of ϕ/ψ torsions, the proteins were reconstructed in PyRosetta and then scored. Figure 4 shows the resulting scores from the 2 custom Rosetta scoring functions, 1 s and 2 s, with the 20 lowest scoring structures in blue. The top 10 solutions for each noise level are depicted in Fig. 5.

Fitting of synthetic data for a transmembrane helical hairpin (pdb id 4a2n)

The first two transmembrane helices of integral membrane methyltransferase protein (PDB id 4a2n) (Yang et al. 2011) were used to calculate the spectrum, totaling 64 residues. The starting orientation, $(\beta, \alpha) = (0.0, 3.14)$, was chosen to orient the helices to be approximately parallel with the membrane normal. All the same noise levels, values of *stride*, and *simits* as for 2gb1 were used for 4a2n. For the three noise levels *tol* was 30 Hz, 50 Hz, 70 Hz, and *tol2* was 5 Hz, 20 Hz, and 20 Hz, respectively. The back-calculated structures have been analyzed and filtered similarly to the GB1 protein, albeit using the scoring functions for membrane proteins. The results are shown in Figs. 7 and 8.

Variations in the tensor orientation

A separate trial was performed to quantify the effects for the potentially variable orientation of the σ_{33} axis of the CSA tensor relative to the NH bond. For this trial, the three-dimensional 2gb1 spectrum was generated by randomly varying the σ_{33} angle off of the NH bond within a uniform distribution from 18° to 19° for every residue; in the previous trials, this angle was set to 18.5° for all residues. The calculated resonances deviated from those calculated by assuming a constant σ_{33} angle by a maximum of 79.7 Hz with a mean of 25.3 Hz per resonance.

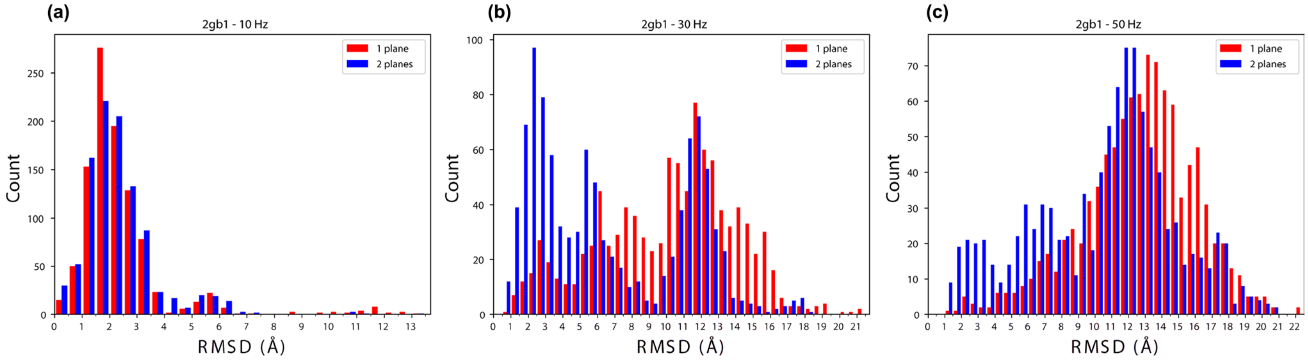


Fig. 3 Structural RMSD histograms (relative to the original PDB structure coordinates measured in Å) for 1000 structures of protein 2gb1 back-calculated from 3 synthetic NMR angular restraints per plane (^{15}N CSA, ^{1}H–^{15}N, ^{1}H$_{\alpha}$–^{13}Cα DC's) with experimental uncertainty of: **a** 10, **b** 30, and **c** 50 Hz. The results of runs involving fitting one plane at a time (*stride* = *1*) are shown in red while the histograms obtained from fitting two planes at a time (*stride* = *2*) are shown in blue. Each histogram contains 1000 runs total for each value of *stride*. Histograms are binned in 0.5 Å steps

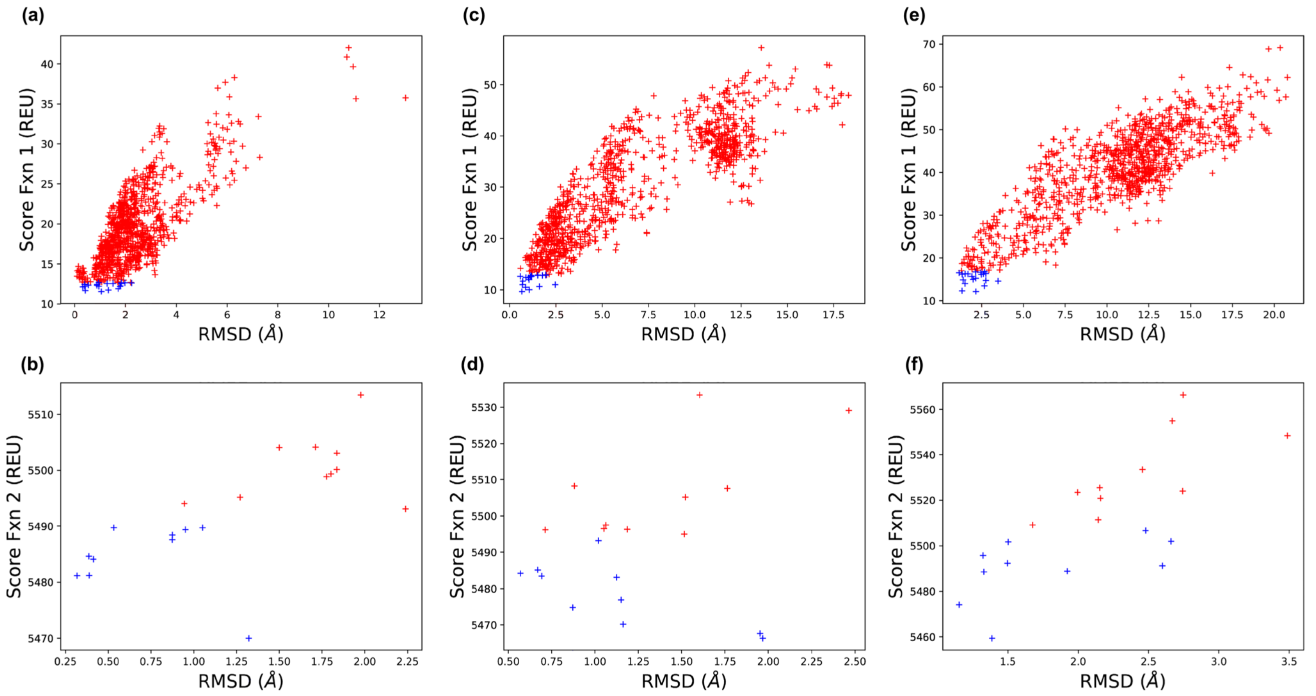
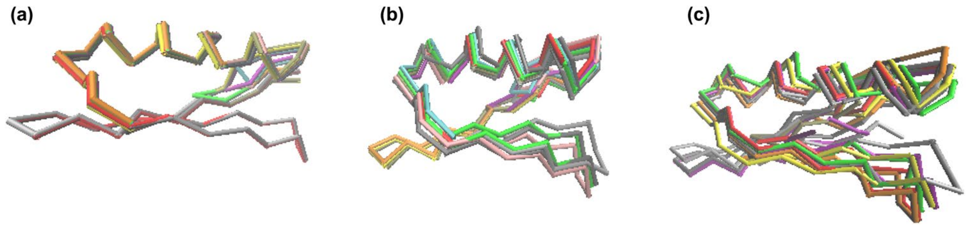


Fig. 4 Scatter plots for structural RMSDs versus Rosetta scores for 2gb1. Plots **a/b**, **c/d** and **e/f**, correspond to noise levels of 10, 30, and 50 Hz, respectively. The y-axes represent the Rosetta function scores, in REU (Rosetta energy units). Top plots result from Score Function 1 s applied to all 1000 structural solutions from the program output. Scatter points in blue are the lowest 20 scoring structures for the Score Function 1 s, which are not necessarily the 20 lowest structural RMSD values. Bottom plots result from Score Function 2 s applied to the 20 lowest scoring structural solutions from the previous scoring function. Scatter points in blue are the 10 lowest scoring structures

Fig. 5 Overlays of top 10 solutions for soluble protein 2gb1 for noise levels **a** 10, **b** 30, and **c** 50 Hz. The 10 solutions for each noise level correspond to the structural RMSD values listed in Table 3



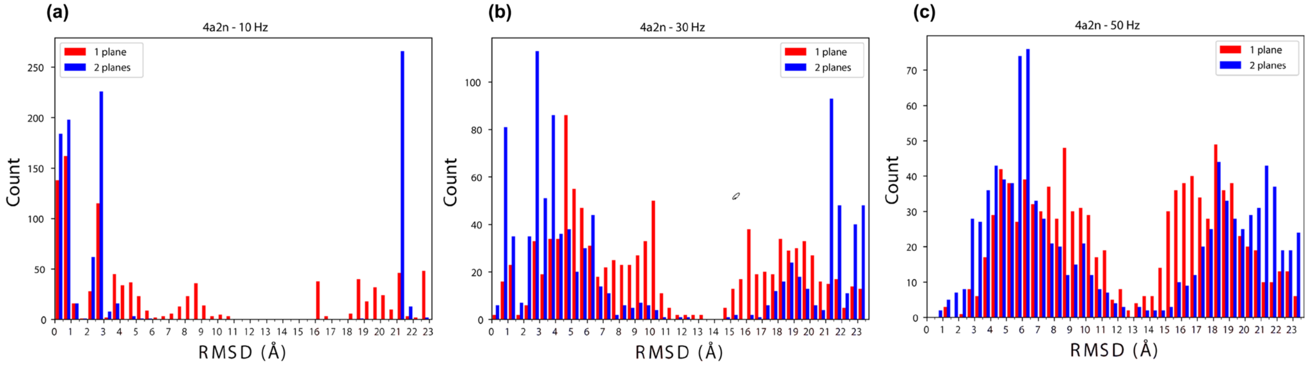


Fig. 6 Structural RMSD histograms (relative to the original structure) for 1000 structures of protein 4a2n back-calculated from 3 synthetic NMR angular restraints per plane (^{15}N CSA, ^{1}H–^{15}N, ^{1}H$_{\alpha}$–^{13}Cα DC's) with experimental uncertainty of: **a** 10, **b** 30, and **c** 50 Hz. The results of runs involving fitting one plane at a time (*stride* = *1*) are shown in red while the histograms obtained from fitting two planes at a time (*stride* = *2*) are shown in blue. Each histogram contains 1000 runs total for each value of *stride*. Histograms are binned in 0.5 Å steps

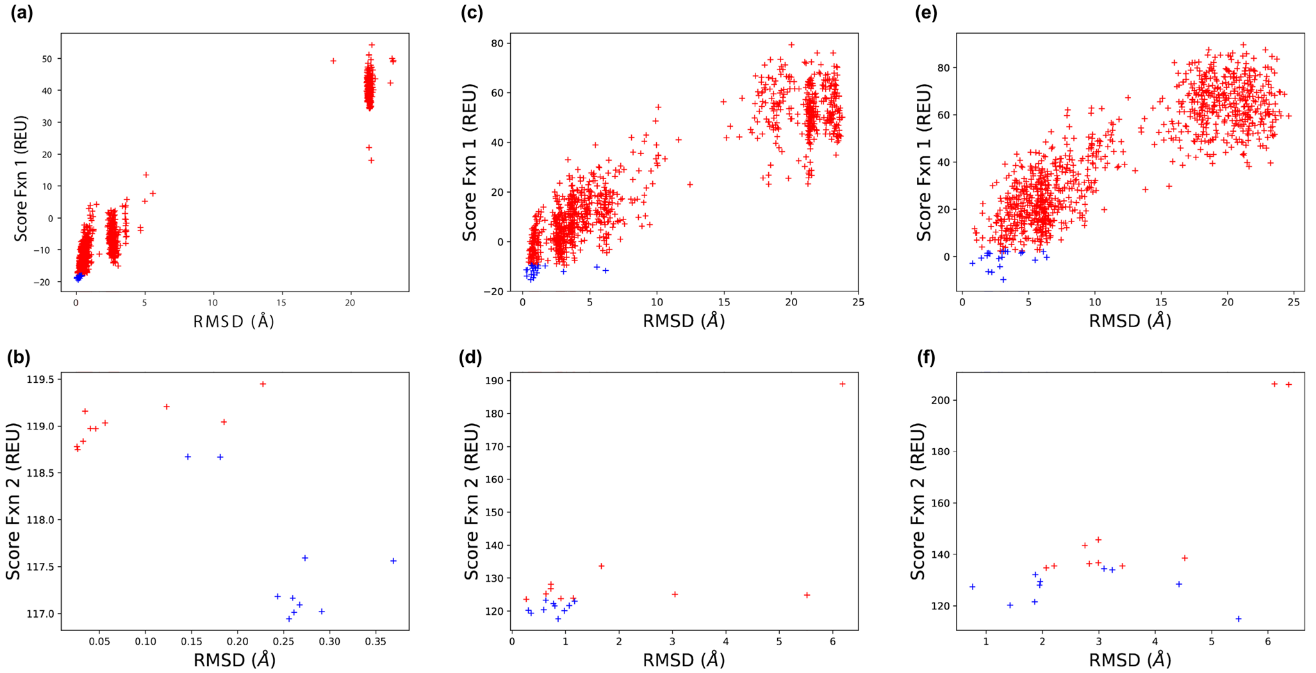


Fig. 7 Scatter plots for structural RMSDs versus Rosetta scores for 4a2n. Plots a/b, c/d. and e/f, correspond to noise levels of 10, 30, and 50 Hz, respectively. The y-axes represent the Rosetta function scores, in REU (Rosetta energy units). Top plots result from Score Function 1 m applied to all 1000 structural solutions from the program output. Scatter points in blue are the lowest 20 scoring structures for the Score Function 1 m, which are not necessarily the 20 lowest structural RMSD values. Bottom plots result from Score Function 2 m applied to the 20 lowest scoring structural solutions from the previous scoring function. Scatter points in blue are the 10 lowest scoring structures

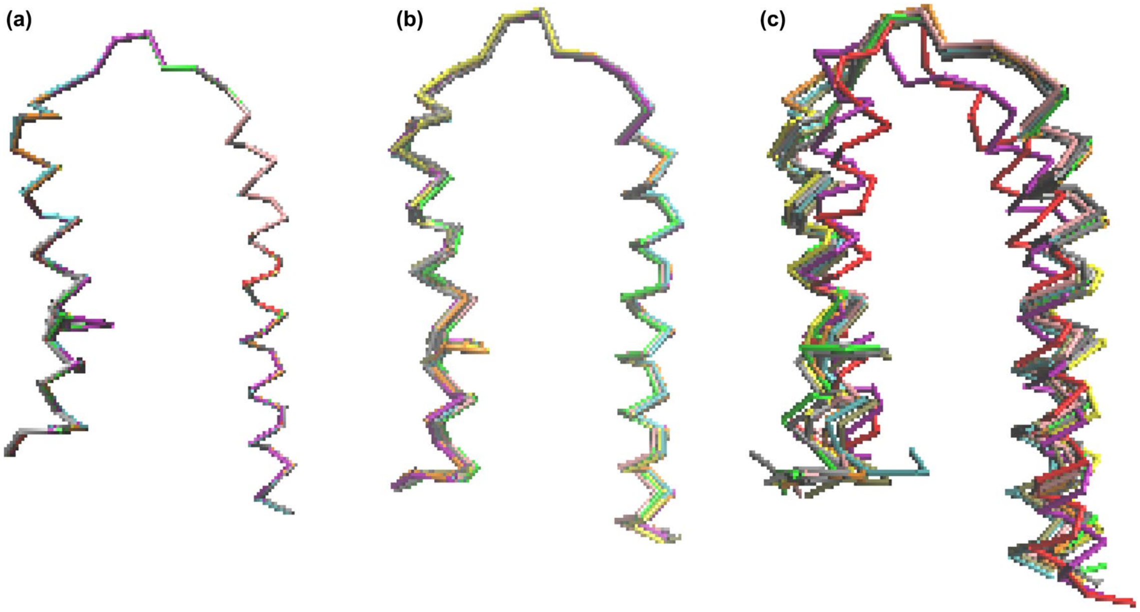


Fig. 8 Overlays of 10 top-scoring structures for membrane protein 4a2n for **a** 10, **b** 30, and **c** 50 Hz of experimental uncertainty, *stride* = 2. The 10 solutions in each picture correspond to the structural RMSD values in Table 3

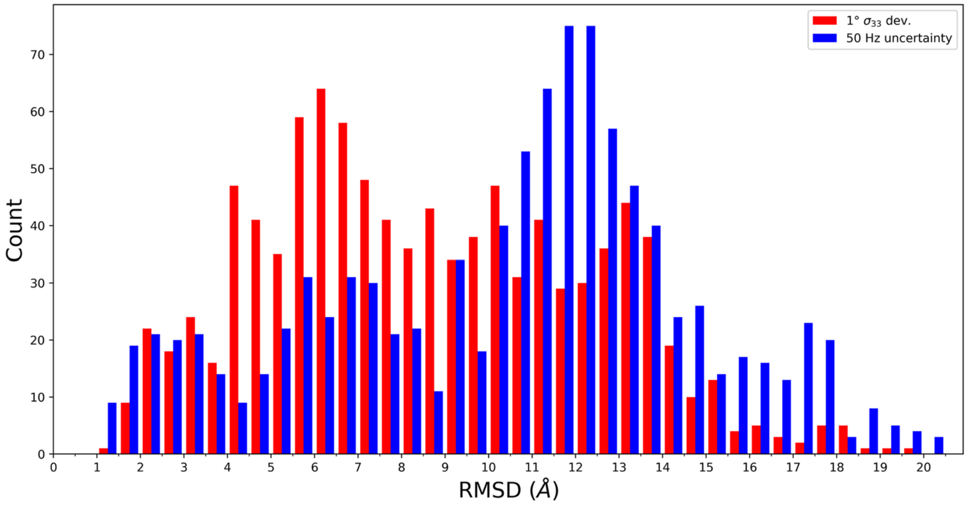


Fig. 9 Structural RMSD histogram for the synthetic spectra of 2gb1 generated by using a random uniform distribution of the σ_{33} angle within 1°, centered at 18.5°. One thousand structures are back calculated on this spectrum using a fixed angle of 18.5° (red). The histogram includes 1000 structures obtained from a spectrum generated by adding random uniform noise of ± 50 Hz (from Fig. 3c) and is shown in blue for direct comparison

Table 2 Spectral RMSDs of fits using one versus two-plane method

Protein	2GB1		4A2N	
Experimental uncertainty (Hz)	1 Plane	2 Planes	1 Plane	2 Planes
10	86	37	165	38
30	543	125	246	119
50	628	212	366	277

All values are in Hz

This mean deviation most closely resembles the trials of Fig. 3c run with 50 Hz of experimental uncertainty,. The structures were then back-calculated by assuming a constant average value for the σ_{33} angle of 18.5°. Figure 9 shows the histogram of the structural RMSD values alongside the results for 2gb1 with 50 Hz of experimental uncertainty for direct comparison (cf. Figure 3c). As can be seen from Fig. 9, the distributions of the structural solutions are similar. Moreover, structural RMSDs well below 2 Å are still obtainable even if the spectra generated with a variable σ_{33} angle are fitted assuming a constant (average) CSA tensor orientation. Thus, the uncertainty in the tensor orientation can be adequately represented by using a uniform experimental uncertainty for the spectral positions.

Discussion

The highly correlated nature of angular restraints results in extreme ruggedness in the spectral RMSD landscape, which may render a global minimization search problematic. While the popular software package Xplor-NIH (Schwieters et al. 2003) predicts structure via molecular dynamics augmented with experimental restraint potentials, such as distance restraints and implicit solvation potentials (Tian et al. 2014; Tian et al. 2017), our algorithm outlined in Fig. 2 attempts a de novo structure prediction from the NMR angular restraints alone. Such a heavy reliance on the angular restraints necessitates at least three (3) NMR restraints per peptide plane in order to obtain a convergent set of

Table 3 Structural RMSD of top 10 solutions for 2gb1 and 4a2n

Uncertainty Level	2gb1			4a2n		
	10 Hz	30 Hz	50 Hz	10 Hz	30 Hz	50 Hz
1	1.32	1.97	1.39	0.26	0.86	5.48
2	0.32	1.95	1.15	0.26	0.35	1.43
3	0.39	1.16	1.33	0.29	0.98	1.86
4	0.41	0.87	1.92	0.27	0.30	0.76
5	0.39	1.15	2.60	0.26	0.59	1.95
6	0.87	1.13	1.50	0.24	0.80	4.43
7	0.87	0.69	1.32	0.37	1.08	1.96
8	0.95	0.57	1.50	0.27	0.77	1.87
9	1.05	0.67	2.66	0.18	1.17	3.24
10	0.53	1.02	2.48	0.15	0.63	3.09

structures. Peptide planes with less than 3 restraints greatly increase the degeneracy of structural solutions, making the correct structure less likely to be found. Due to the above reasons, a sequential walk along the backbone appears at present to be a more practical method for solving a structure entirely from NMR angular restraints. Using the simplex solver results in the correct torsion angles for the majority of residues, even when 50 Hz of experimental uncertainty is present throughout the spectrum. Yet miscalculating only a few pairs of the torsion angles, especially those in turn regions between the secondary structure elements, may result in the calculated structures being very different from the real structure. Even though some segments of the calculated structure may closely match parts of the real structure, the overall tertiary fold can still have a high structural RMSD value. Despite this potential pitfall our algorithm is capable of sampling a vast conformational space thereby retaining the relevant torsion angles along the backbone. In some cases, occasionally missed torsional pairs along the backbone can be mitigated by selecting compensatory ϕ/ψ torsions in the subsequent residues that would largely return the subsequent plane(s) to the correct orientation.

A notable modification of the algorithm as compared to the previous work (Yin and Nevzorov 2011) involves the capability of fitting NMR resonances simultaneously for two peptide planes (4 torsion angles) instead of fitting a single plane at a time. As previously mentioned, a potential pitfall for any sequential structure fitting algorithm is that only a few incorrectly determined torsion angles may compromise the overall fold and yield high structural RMSDs to the actual structure. This can occur even if the fit to the spectrum is acceptable. A plausible reason for the discord between the spectral RMSD versus structural RMSD is that incorrect torsions can provide better fits to the spectrum for certain residues (especially if experimental uncertainty is large), thus being more likely selected by the minimization algorithm. Adding more residues to be fit simultaneously helps in rejecting the (incorrect) solutions that may fit the i'th residue very closely while being inconsistent with the NMR data for the residues downstream (i.e. $i+1$ and $i+2$) due to the incompatible orientations of their peptide planes. It was determined that increasing the number of simultaneously fitted planes in the simplex solver to just two largely avoids over-fitting certain NMR resonances while being computationally feasible. The solver could always find dozens of structural solutions within 2 Å structural RMSD relative to the starting structure (among 1000 iterations). Adding a third plane to the solver required an order of magnitude more simplex iterations to reliably find the correct solution, which considerably slowed down the execution time while not appreciably affecting the efficacy of the search.

A direct comparison between the above two methods of fitting is presented in Table 2 and Figs. 3 and 6. With the (minor) exception of 4a2n at 50 Hz experimental error, the spectral RMSDs (in Hz) for the one-plane method are more than double that for the 2 planes. Since these values are averages of 1000 structures in each case, it can be concluded that multiple planes consistently improve the fit to the spectrum. More importantly, this should lead to solutions with lower structural RMSDs to the PDB structures. Although there is no guarantee that a single structural solution that better fits the spectrum will have a lower structural RMSD to the PDB coordinates, on average the distribution of structural RMSDs shifts towards lower values as compared to under-fitted spectra. This can be seen in the histograms of Figs. 3 and 6 from the intensity levels on the left hand side (< 2 Å RMSD). Owing to their low structural RMSDs, the fraction of these solutions relative to the total number of possible structural solutions determines the success rate for the algorithm. At every noise level there is consistently a higher proportion of sub-2 Å solutions for the two-plane fitting than for the single plane fitting among the 1000 calculated structures. While the improvement is marginal at 10 Hz uncertainty level for 2gb1 (cf. Figure 3a), fitting two planes at a time has proven to have a more pronounced effect when the experimental uncertainty is increased (Figs. 3b, c), where the ratio of low-RMSD solutions for two- versus single-plane fitting exceeds 5:1. The same trend persists for 4a2n (Fig. 5): at 50 Hz of experimental uncertainty there are roughly twice as many "correct" solutions for 4a2n when two peptide planes are fit instead of one. These results suggest that fitting two planes simultaneously in the simplex solver is highly beneficial for retaining good structural solutions without considerably increasing the program runtime.

Nevertheless, the histograms in Figs. 3 and 6 show that even at the lowest noise levels, the majority of the solutions are still expected to have greater than 2 Å structural RMSD, thus being unacceptable. At the highest noise levels, less than 1% of the solutions contain acceptable structures. Since all these solutions fit the CSA and dipolar couplings almost equally well, it is clear that fitting the angular restraints cannot by itself ensure correctness of the calculated structure, except, perhaps single helices with highly constrained torsion angles (Thiriot et al. 2004). The obtained 1000 structural solutions would still likely contain a number of acceptable structures, at which point the problem becomes to distinguish the realistic structural solutions from the implausible ones. Through various Rosetta scoring terms, which are compiled from a plethora of statistics derived from real structures, one can identify the most likely global folds amongst the structures calculated from the spectral fit.

The scatter plots of Fig. 4 demonstrate that the custom Rosetta scoring functions for 2gb1 are capable of filtering out the most implausible solutions. Although some low RMSD structures may have a higher/worse Rosetta score, it is important to note that the inverse generally does not happen, i.e. high-RMSD structures do not yield lowest/best Rosetta scores. For Score Function 1 s the overall correlations between structural RMSD and score are strong, but these correlations alone do not constitute a sole reliable metric for obtaining the final results. The primary goal for the first scoring function is to separate out the most unrealistic solutions, while retaining low to moderate RMSD structures. Therefore, the first selection of solutions (20 for these calculations) may need to be subjected to the second scoring function. The distribution of structural RMSDs in the top 20 solutions can be seen in Figs. 4b, d, f. For each noise level the Score Function 1 s solutions contained structures over 2 Å. Score Function 2 s appears to have a stronger correlation between the structural RMSD (in Å) and Rosetta score, and filters out the remaining unrealistic structures among these 20 solutions from the realistic, low RMSD ones. The results for 2gb1 in Table 3 shows that applying different custom Rosetta scoring functions in series can be very effective at all noise levels, with the majority of structures selected in the final list having very low structural RMSD values. As the experimental uncertainty increases, a number of higher-RMSD structures may still end up among the final 10 solutions. Neither 10 nor 30 Hz experimental uncertainty contains any structures having RMSDs greater than 2 Å, but for 50 Hz there are three such structures (in between 2 and 3 Å).

The results for 4a2n are very similar to that of 2gb1: the post-fitting Rosetta protocol yielded 10 solutions with largely acceptable structural RMSDs at all experimental uncertainty levels. One noticeable difference is that the final top scoring solution for 50 Hz has a significantly higher RMSD of 5.48 Å (cf. Table 3); additionally there were three other structures present having RMSD > 2 Å. Closer observation of Fig. 7f shows that a number of lower RMSD structures, between 2 and 3.5 Å, were spurned in favor of a few high RMSD structures with low score function 2 scores. Overall, most of the structures from 50 Hz were acceptable, yielding a consensus structure. Another feature of the 4a2n trials that distinguishes it from that of 2gb1 is that the RMSDs of the structural solutions are distributed less continuously. Figures 6a, c, e show that Rosetta segregates the structural solutions essentially into two groups with higher and lower RMSD values. A possible reason for this segregation is that 4a2n could be considered structurally simpler than 2gb1. In general, 4a2n has 2 secondary structural elements (α-helices) separated by a single turn, whereas 2gb1 has 1 helix and 2 β-sheets that are separated by 3 turn regions overall. It is likely that the turn regions pose consequential obstacles for the algorithm, in that they determine the relative orientations of the larger secondary structure elements. Nevertheless, the post-fitting Rosetta screening protocol still correctly identified the low-RMSD structures amongst the 1000 total solutions.

In general, the utilization of the *mp_env* and *rama* terms are especially useful for screening the calculated structures of membrane proteins. Frequently, in solving for 4a2n structure, the algorithm would fail to orient the second helix back into the membrane, yielding one long, straight helix. The term *mp_env* describes the burial depth into the membrane and thus heavily penalizes such a structure where a hydrophobic α-helix is exposed to a soluble environment. The *rama* term (and similar term *p_aa_pp*) is also indispensible in filtering out physically impossible structures and providing a simultaneous validation for the calculated torsion angles to comply with the Ramachandran map. Finally, in this study it was shown that two scoring functions were sufficient to achieve the end goal of separating low RMSD structures from the high ones. It is possible that additional scoring functions in the protocol may be necessary to achieve better scoring of the structural solutions calculated from NMR angular restraints.

Conclusions

A critical analysis for the feasibility of *de-novo* structure calculations using NMR angular restraints in uniaxially aligned samples has been presented. The sequential-fitting structure calculation algorithm has been substantially improved to better handle experimental uncertainty. By doubling the number of ϕ/ψ torsion angles to be fitted at once, the output structural solutions have been more effectively funneled toward lower RMSD structures. The post-fitting filtering was made possible by using several custom Rosetta scoring functions including statistical and physical terms that correlate well with low structural RMSDs relative to the original protein structures. To test the robustness of the structure determination protocol, synthetic NMR datasets derived from two proteins of different topology were employed. A set of 1000 calculated solutions has been reduced down to 10 acceptable solutions, of which the majority of structural RMSD scores were less than 2 Å. Thus, without knowing the true structure a priori, it is possible to first calculate a large set of the structural solutions consistent with the NMR angular restraints and then filter them using the Rosetta scoring functions to obtain the most plausible structure(s).

Experimental error randomly added to the synthetic spectral resonances can adequately account for the deviations from constant CSA tensor orientation. Presumably, adding such random error can also encompass non-ideal peptide plane geometry, variability in the CSA tensor principal values, and insufficient spectral resolution. About 50 Hz of experimental error (or 1 ppm at 50 MHz ^{15}N NMR frequency) likely represents the maximum degree of uncertainty that could be tolerated in three-dimensional experiments to calculate structures reliably by using the method of ssNMR angular restraints and Rosetta-based filtering. For both protein structures considered, the maximum 50 Hz uncertainty yielded less than 3% of the 1000 total solutions that had a structural RMSD < 2 Å. Greater experimental uncertainty in angular restraints and missing data would yield a highly diminishing proportion of acceptable solutions, which may necessitate adding a fourth dimension to the NMR spectrum, such as ^{13}C–^{15}N dipolar couplings to increase the likelihood of obtaining correct protein folds. Finally, the presented framework for de-novo structure determination of protein structures from NMR angular restraints calls for the necessity of the development of pulse sequences correlating ^{15}N spins with $^{1}H_{\alpha}$–$^{13}C_{\alpha}$ and ^{13}C–^{15}N dipolar couplings for macroscopically aligned membrane proteins.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. 1508400 and in part by the Army Research Office Grant W911NF1810363.

References

Alford RF, Leman JK, Weitzner BD, Duran AM, Tilley DC, Elazar A, Gray JJ (2015) An integrated framework advancing membrane protein modeling and design. PLoS Comput Biol 11:1–23

Bertram R, Asbury T, Fabiola F, Quine JR, Cross TA, Chapman MS (2003) Atomic refinement with correlated solid-state NMR restraints. JMR 163:300–309

Bryson M, Tian F, Prestegard JH, Valafar H (2008) REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data. JMR 191:322–334

Chaudhury S, Lyskov S, Gray JJ (2010) PyRosetta: a script-based interface for implementing molecular modeling algorithm using Rosetta. Bioinformatics 26:689–691

Chellapa GD, Rose GD (2015) On interpretation of protein X-ray structures: planarity of the peptide unit. Proteins 83:1687–1692

Cornilescu G, Bax A (2000) Measurement of proton, nitrogen, and carbonyl chemical shielding anisotropies in a protein dissolved in a dilute liquid crystalline phase. J Am Chem Soc 122:10143–10154

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:211–222

Dvinskikh S, Yamamoto K, Ramamoorthy A (2006) Heteronuclear isotropic mixing separated local field NMR spectroscopy. J. Chem. Phys. 125:034507

Gayen A, Banigan JR, Traaseth NJ (2013) Ligand-Induced Conformational changes of the multidrug resistance transporter EmrE probed by oriented solid-state NMR spectroscopy. Angew Chem Int Ed 52:10321–10324

Gleason NJ, Vostrikov VV, Greathouse DV, Koeppe RE (2013) Buried lysine, but not arginine, titrates and alters transmembrane helix tilt. Proc Natl Acad Sci 110:1692–1695

Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM (1991) A novel highly stable fold of the immunoglobin binding domain of streptococcal protein G. Science 253:657–661

Herrmann T, Guntert P (2002) Protein NMR structure determination and automated NOE assignment using the new software CANDID and the torsion angle dynamics alogrithm DYANA. JMB 319:209–227

Kabsch W (1976) A solution for the best rotation to relate two sets of vectors. Acta Crystallogr A A32:922–923

Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97(19):10383–10388

Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302(5649):1364–1368

Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, Gray JJ, Kortemme T, Richardson JS, Havranek JJ, Snoeyink J, Baker D, Kuhlman B (2013) Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol 523:109

Lee DK, Wittebort RJ, Ramamoorthy A (1998) Characterization of ^{15}N chemical shift and ^{1}H–^{15}N dipolar coupling interactions in a peptide bond of uniaxially oriented and polycrystalline samples by one-dimensional dipolar chemical shift solid-state NMR spectroscopy. JACS 120:8868–8874

Leman JK, Ulmschneider MB, Gray JJ (2014) Computational modeling of membrane proteins. Proteins Struct Funct Bioinf 83:1–24

Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19:315–316

Marassi FM, Opella SJ (2003) Simultaneous assignment and structure determination of a membrane protein from NMR orientational restraints. Protein Sci 12:403–411

McDonnell PA, Shon K, Kim Y, Opella SJ (1993) fd Coat protein structure in membrane environments. JMR 233:447–463

Nevzorov AA, Opella SJ (2007) Selective averaging for high-resolution solid-state NMR spectroscopy of aligned samples. JMR 185:59–70

O'Meara MJ, Leaver-Fay A, Tyka M, Stein A, Houlihan K, DiMaio F, Bradley P, Kortemme T, Baker D, Snoeyink J (2015) A Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta. J Chem Theory Comput 11:609–622

Opella SJ, Marassi FM, Gesell JJ, Valente AP, Kim Y, Oblatt-Montal M, Montal M (1999) Structures of the M2 channel-lining segments from the nicotinic acetylcholine and NMDA receptors by NMR spectroscopy. Nat Struct Mol Biol 6:374

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

Ruan K, Briggman KB, Tolman JR (2008) De novo determination of internuclear vector orientations from residual dipolar coupling measured in three independent alignment media. J Biomol NMR 41:61–76

Saito H, Ando I, Ramamoorthy A (2010) Chemical shift tensor—the heart of NMR: insights into biological aspects of proteins. Prog Nucl Magn Reson Spectrosc 57:181–228

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. JMR 160:65–73

Sharma M, Yi M, Dong H, Qin H, Peterson E, Busath DD, Cross TA (2010) Insight into the mechanism of the influenza A proton channel from a structure ina lipid bilayer. Science 330:509–512

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS^{+}: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Sinha N, Grant CV, Park SH, Brown JM, Opella SJ (2007) Triple resonance experiments for aligned sample solid-state NMR of 13C and 15N labeled proteins. J Magn Reson 186:51–64

Spera S, Bax A (1991) Empirical correlation between protein backbone conformation and Ca and Q8 13C nuclear magnetic resonance chemical shifts. J Am Chem Soc 113:5490–5492

Stewart PL, Valentine KG, Opella SJ (1987) Structural analysis of solid-state NMR measurements of peptides and proteins. JMR 71:45–61

Stewart PL, Tycko R, Opella SJ (1988) Peptide backbone conformation by solid-state nuclear magnetic resonance spectroscopy. J Chem Soc 84:3803–3819

Thiriot DS, Nevzorov AA, Opella SJ (2004) Structural basis of the temperature transition of Pf1 bacteriophage. Protein Sci 14:1064–1070

Tian Y, Schwieters CD, Opella SJ, Marassi FM (2014) A practical implicit solvent potential for NMR structure calculation. J Magn Reson 243:54–64

Tian Y, Schwieters CD, Opella SJ, Marassi FM (2017) High quality NMR structures: a new force field with implicit water and membrane solvation for Xplor-NIH. J Biomol NMR 67:35–49

Traaseth NJ, Buffy JJ, Zamoon J, Veglia G (2006) Structural dynamics and topology of phospholamban in oriented lipid bilayers using multidimensional solid-state NMR. Biochemistry 45:13827–13834

Traaseth NJ, Shi L, Verardi R, Mullen DG, Barany G, Veglia G (2009) Structure and topology of monomeric phospholamban in lipid membranes determined by a hybrid solution and solid-state NMR approach. Proc Natl Acad Sci 106:10165–10170

Valafar H, Prestegard JH (2004) REDCAT: a residual dipolar coupling analysis tool. JMR 167:228–241

Verardi R, Shi L, Traaseth NJ, Walsh N, Veglia G (2011) Structural topology of phospholamban pentamer in lipid bilayers by a hybrid solution and solid-state NMR method. Proc Natl Acad Sci 108:9101–9106

Wang J, Kim S, Kovacs F, Cross TA (2001) Structure of the transmembrane region of the M2 protein H+ channel. Protein Sci 10:2241–2250

Wu CH, Ramamoorthy A, Opella SJ (1994) High-resolution heteronuclear dipolar solid-state NMR spectroscopy. JMR 109:270–272

Yamamoto K, Durr UH, Xu J, Im SC, Waskell L, Ramamoorthy A (2013) Dynamic interaction between membrane-bound full-length cytochrome P450 and cytochrome b 5 observed by solid-state NMR spectroscopy. Scientific Rep 3:2538

Yang J, Kulkarni K, Manolaridis I, Zhang Z, Dodd RB, Mas-Droux C, Barford D (2011) Mechanism of isoprenylcysteine carboxyl methylation from the crystal structure of the integral membrane methyltransferase ICMT. Mol Cell 44:997–1004

Yarov-Yarovoy V, Schonbrun J, Baker D (2005) Multipass membrane protein structure prediction using Rosetta. Prot Struct Funct Bioinf 62:1010–1025

Yin Y, Nevzorov AA (2011) Structure determination in “shiftless” solid state NMR of oriented protein samples. JMR 212:64–73